Patrona, F., Iosifidis, A., Tefas, A., Nikolaidis, N., & Pitas, I. (2016). Visual Voice Activity Detection in the Wild. *IEEE Transactions on Multimedia*, *18*(6), 967-977. https://doi.org/10.1109/TMM.2016.2535357

Visual Voice Activity Detection in the Wild

Foteini Patrona*, Alexandros Iosifidis†, Anastasios Tefas*, Nikolaos Nikolaidis*, and Ioannis Pitas*
*Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece.
{tefas,nikolaid,pitas}@aiia.csd.auth.gr.
†Department of Signal Processing, Tampere University of Technology, Tampere, Finland
{aiosif}@aiia.csd.auth.gr.

***Abstract*—The Visual Voice Activity Detection (V-VAD) problem in unconstrained environments is investigated in this paper. A novel method for V-VAD in the wild, exploiting local shape and motion information appearing at spatiotemporal locations of interest for facial video description and the Bag of Words (BoW) model for facial video representation, is proposed. Facial video classification is subsequently performed using state-of-the-art classification algorithms. Experimental results on one publicly available V-VAD data set denote the effectiveness of the proposed method, since it achieves better generalization performance in unseen users, when compared with recently proposed state-of-the-art methods. Additional results on a new, unconstrained, data set provide evidence that the proposed method can be effective even in such cases in which any other existing method fails.**

***Index Terms*—Voice Activity Detection in the wild, Space-Time Interest Points, Bag of Words model, kernel Extreme Learning Machine, Action Recognition**

I. Introduction

The task of identifying silent (vocal inactive) and non-silent (vocal active) periods in speech, called Voice Activity Detection (VAD) has been widely studied for many decades using audio signals. In the last two decades, though, considerable attention has been paid to the use of visual information, mainly as an aid to the traditional Audio-only Voice Activity Detection (A-VAD), due to the fact that, contrary to audio, visual information is insensitive to environmental noise and can, thus, be of help to A-VAD methods for speech enhancement and recognition [1], speaker detection [2], segregation [3] and identification [4] as well as speech source separation [5], [6] in noisy and reverberant conditions or in Human Computer Interfaces (HCIs).

All V-VAD methods proposed in the literature till now set several assumptions concerning the visual data recording conditions, which are rather constraining in their vast majority. In brief, the available data sets used for evaluating the performance of such methods are recorded indoors, under fully constraint conditions, e.g., using preset static illumination, simple background and no or negligible background noise produced by humans speaking or by other sound sources. Moreover, no or slight speaker movements are encountered and the recording setting is calibrated so that the entire speaker face as well as the mouth are always fully visible from a camera positioned right in front of the speaker, so that special features describing their shape and/or motion can be calculated. That is, the human face should have a frontal orientation with respect to the capturing camera and the facial Region Of Interest (ROI) should have adequate resolution (in pixels). Such a scenario restricts the applications, where V-VAD methods can be exploited. For example, in movie (post-)production, the persons/actors are free to move and their facial pose may change over time, as is also the case in all the places where audio-visual surveillance would be of interest. Most V-VAD methods proposed in the literature would probably fail in such an application scenario. Last but not least, most currently existing methods focus on the accurate detection of the visually silent intervals in a video sequence, which in general is not as challenging as the accurate detection of the visually speaking intervals, due to the fact that the latter can be easily confused with intervals of laughter, mastication or other facial activities. The aforementioned difficulty of distinguishing especially between laughter and speech is highlighted in [7], where a method exploiting both audio and visual information aiming at an effective discrimination is presented.

Non-invasive V-VAD, where the persons under investigation are free to change their orientation and their distance from the capturing camera, is within the scope of this paper. Inspired by relative research in human action recognition [8], [9], [10], this unconstrained V-VAD problem will subsequently be mentioned as *V-VAD in the wild*. While human action recognition in the wild has been extensively studied in the last decade and numerous methods addressing this problem have been proposed, V-VAD in the unconstrained case has not been addressed yet. In this paper, a method oriented at dealing with the problem of V-VAD in the wild is proposed, having as only prerequisite assumption that the faces appearing in the facial moving region videos being processed can be automatically detected using a face detection algorithm and tracked for a number of consecutive frames.

The proposed method is formed by three processing steps. In the first step, a face detection technique [11] is applied to a video frame, in order to determine the facial Region of Interest (ROI), which is subsequently tracked over time [12], in order for a facial ROI trajectory of the person under investigation to be created. Such videos are noted as *facial moving regions* hereafter. In the second step, local shape and motion information appearing in spatiotemporal video locations of interest is exploited for the facial moving region video representation. To this end, two facial moving region representation approaches are evaluted, a) Histogram of Oriented Gradient (HOG) and Histogram of Optical Flow (HOF) descriptors calculated on Space Time Interest Point

(STIP) video locations [8] and b) HOG, HOF and Motion Boundary Histogram (MBHx, MBHy) descriptors calculated on the trajectories of video frame interest points that are tracked for a number of L consecutive frames [9]. Both facial moving region descriptors are combined with the Bag of Words (BoWs) model [13], [14], in order to determine facial moving region video representations.

Finally, facial moving region video classification in visually silent and visually speaking ones is performed employing a Single Hidden Layer Feedforward Neural (SLFN) network, trained by applying the recently proposed kernel Extreme Learning Machine (kELM) classifier [15], [16]. A facial moving region verification step is introduced before this step, in cases where videos not depicting facial images may be encountered, in order to ensure that only facial moving region videos are going to be classified as visually silent and non-silent, by performing facial moving region - non facial moving region video classification. The proposed approach is evaluated on a publicly available V-VAD data set, namely CUAVE [17], where it is shown to outperform recently proposed V-VAD methods to a large extend. In addition, a new V-VAD data set, extracted from full length movies in order to evaluate the performance of the proposed approach on a case of V-VAD in the wild was created. Experimental results denote that the proposed approach can operate reasonably well in the cases where other V-VAD methods fail.

The remainder of this paper is organized as follows. Section II discusses previous work on V-VAD. The proposed V-VAD approach is described in Section III. The data sets used in our experiments and the respective experimental results are presented in Section IV. Finally, conclusions are drawn in Section V.

II. Previous Work

V-VAD methods proposed in the literature can be roughly divided in model-based and model-free ones. Model-based methods require a training process, where positive and negative paradigms are employed for model learning. In model-free methods, no direct training is performed, thus circumventing the need for an a-priori knowledge of the data classes at the decision stage. Moreover, either visual only or audiovisual data features can be exploited. In the latter case, combination of the audio and video modalities can be achieved in two different ways, either by combining the audio and visual features (feature/early fusion) or by performing A-VAD and V-VAD independently and fusing the obtained classification results (decision/late fusion) [18].

Model-free V-VAD methods, usually rely solely on combinations of speaker-specific static and dynamic visual data parameters, like lip contour geometry and motion [19], or inner lip height and width trajectories [20] that are compared to appropriate thresholds for decision making. Emphasis is given on dynamic parameters due to the fact that identical lip shapes can be encountered both in silent and non-silent frames, making static features untrustworthy. In both these approaches, there is no discrimination between speech and non-speech acoustic events, which are thus handled as non-silent sections. Another model-free approach is proposed in [21], where signal detection algorithms are applied on mouth region pixel intensities along with their variations, in order to discriminate between speech and non-speech frames.

Concerning model-based V-VADs, features like lip opening, rounding and labio-dental touch (a binary feature indicating whether the lower lip is touching the upper teeth) for lip configuration followed by motion detection and SVM classification are proposed in [22], in an attempt to distinguish between moving and non-moving lips and then between lip motion originating either from speech or from other face/mouth activities, e.g., from facial expressions or mastication [19], [20]. Such a VAD system can constitute the first stage of a Visual Speech Recognition (VSR) system. The discriminative power of static and dynamic visual features in V-VAD is investigated in [23], where the predominance of dynamic ones is highlighted. The same approach is also adopted in [24], where facial profile as well as frontal views are used. Though not providing as much useful information as the frontal ones, facial profile views are proven to be useful in VAD. A greedy snake algorithm exploiting rotational template matching, shape energy constraints and area energy for lip extraction avoiding common problems resulting from head rotation, low image resolution and active contour mismatches is introduced in [25], where adaboost is used for classifier training. Adaboost is also used in [5] for the V-VAD classifier training, of a system performing Blind Source Separation (BSS) based on interference removal, after the extraction of lip region geometric features. Finally, HMMs are used in [26] to model the variation of the optical flow vectors from a speaker mouth region during non-speech periods of mouth activity.

An early-fusion model-based AV-VAD approach is introduced in [27]. 2D discrete cosine transformations (2D-DCTs) are extracted from the visual signal and a pair of GMMs is used for classification of the feature vector. V-VAD accuracy is quite high in the speaker-dependent case. However, it dramatically decreases in the speaker-independent case experiments, conducted on a simplistic dataset called GRID [28]. Color information is used in the V-VAD subsystem proposed in [29] for skin and lip detection, followed by video-based HMMs aiming to distinguish speech from silence, while lip optical flow input provided to SVMs is employed in [6] for utilization of the visual information, subsequently combined with audio information for multispeaker mid-fusion AV-VAD and Sound Source Localization (SSL).

III. Proposed V-VAD method

The proposed method operates on grayscale facial moving regions. Face detection and tracking [11], [12] techniques are used to find such regions in a video. After determining the facial Regions of Interest (ROIs) in each facial video sequence, we find the union $\mathcal{R} = \{\cup \mathcal{R}_k,\, k = 1, \ldots, K\}$ of all ROIs $\mathcal{R}_k$ within this video sequence. Then, we use this new ROI $\mathcal{R}$ for positioning the face in each video frame and we resize it to a fixed size of $H \times W$ pixels in order to produce the so called *facial video segments*. Subsequently, we apply the proposed V-VAD method. In this Section, we describe each step of the proposed V-VAD method in detail.

A. STIP-based facial video representation

Let $\mathcal{U}$ be an annotated facial video segment database containing N facial videos, which are automatically preprocessed, in order to determine the relevant set of STIPs. In this paper, the Harris3D detector [30], which is a spatiotemporal extension of the Harris detector [31] is employed, in order to detect spatiotemporal video locations, where the image intensity values undergo significant spatiotemporal changes. After STIP localization, each facial video is described in terms of local shape and motion by a set HOG/HOF descriptors (concatenation of L_2 normalized HOG and HOF descriptors) $\mathbf{p}_{ij}$, $i = 1, \ldots, N$, $j = 1, \ldots, N_i$, where i refers to the facial video index and j indicates the STIP index detected in facial video i. In the conducted experiments, the publicly available implementation in [32] has been used for the calculation of HOG/HOF descriptors. An example of STIP locations on

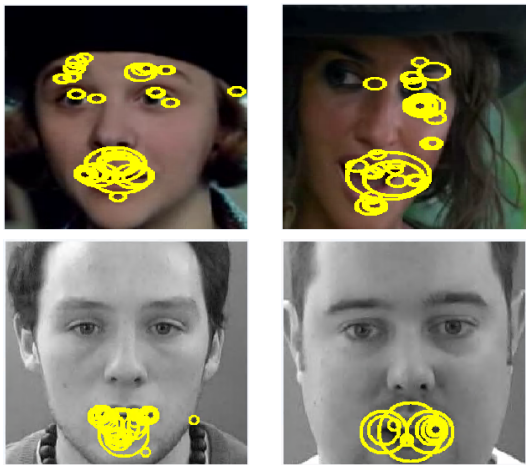

Fig. 1. Examples of detected STIPs on facial videos.

facial videos is illustrated in Figure 1. $\mathbf{p}_{ij}$, $i = 1, \ldots, N$, $j = 1, \ldots, N_i$ are clustered by applying K-Means [33] and the cluster centers $\mathbf{v}_k$, $k = 1, \ldots, K$ form the so-called codebook, i.e., $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$. The descriptors $\mathbf{p}_{ij}, j = 1, \ldots, N_i$ are subsequently quantized using $\mathbf{V}$ and l_1 normalized in order to determine the BoW-based video representation of facial video i, $\mathbf{s}_i \in \mathbb{R}^K$. $\mathbf{s}_i$ are noted as *facial motion vectors* hereafter.

B. Dense Trajectory-based facial video representation

In Dense Trajectory-based facial video segment description [9], video frame interest points are detected on each video frame and are tracked for a number of L consecutive frames. Subsequently, $D = 5$ descriptors, i.e., HOG, HOF, MBHx, MBHy and the (normalized) trajectory coordinates, are calculated along the trajectory of each video frame point of interest. The publicly available implementation in [9] for the calculation of the Dense Trajectory-based video description was used in the conducted experiments. An example of Dense Trajectory locations on facial videos is illustrated in Figure 2. Let us denote by $\mathbf{s}_{ij}^d$, $i = 1, \ldots, N$, $j = 1, \ldots, N_i$, $d = 1, \ldots, D$ the set of descriptors calculated for the N facial video segments in $\mathcal{U}$. Five codebooks $\mathbf{V}_d, d = 1, \ldots, D$ are obtained by applying K-Means on $\mathbf{s}_{ij}^d$ for the determination of K prototypes for each descriptor type. The descriptors $\mathbf{s}_{ij}^d$, $j = 1, \ldots, N_i$ are subsequently quantized using $\mathbf{V}_d$ in order to determine D BoW-based video representations for facial video i.

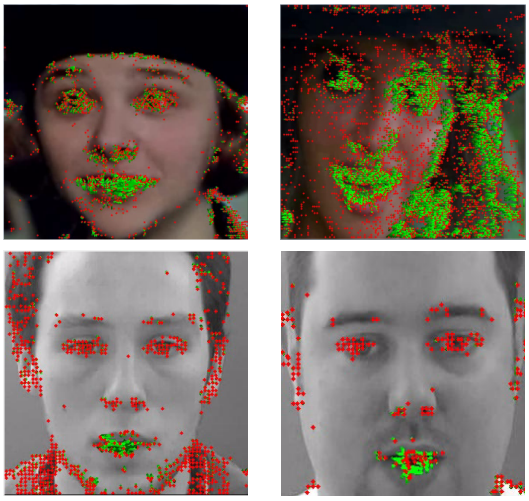

Fig. 2. Examples of Dense Trajectories on facial videos.

C. Facial video segment verification

Due to the fact that the proposed method aims to be applicable in the wild, and on real life recordings, it would be rather inaccurate and optimistic to consider that the face detection and tracking algorithms [11], [12] applied, perform flawlessly and, thus, only facial video segments are produced. For this reason, and in order for a fully automatic approach, not requiring human intervention, to be proposed, a facial video segment verification step had to be introduced before the facial video segment classification as visually silent and visually speaking. In this step, videos are being indeed facial videos or not. Both the STIP and the Dense Trajectory-based video representations are employed in this step, and thus, when a test video is introduced to the pretrained SVM or the SLFN network, the corresponding descriptors are calculated on the video locations of interest and transformed to feature vectors, which are subsequently quantized with the aid of the codebook vectors, in order to produce the facial vector and introduce it to the trained classifiers. Based on the obtained responses, the video is classified as being a facial video segment or not, and the videos identified as non-facial moving regions are discarded from the data set, thus not introduced to the second layer of classifiers, performing V-VAD.

D. SLFN classification

After the calculation of the facial vectors $\mathbf{s}_i \in \mathbb{R}^K$, $i = 1, \ldots, N$ obtained by using the STIP or the Dense Trajectory-based facial video representation, they are used to train a SLFN network. Since both face verification and V-VAD correspond to two-class problems, the network should consist of K input, L hidden and one output neurons, as illustrated in Figure 3. The number L of hidden layer neurons is, usually, much greater than the number of classes involved in the classification problem [10], [15], i.e., $L \gg 2$.

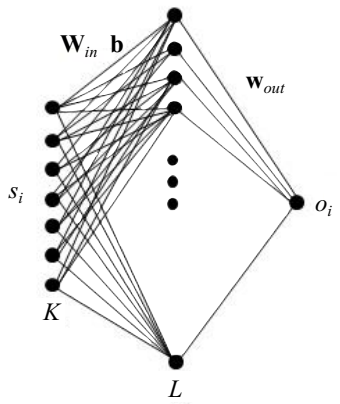

Fig. 3. SLFN network topology for V-VAD.

The network target values t_i, $i = 1, \ldots, N$, each corresponding to a facial vector $\mathbf{s}_i$, are set to $t_i = 1$ or $t_i = -1$, depending on whether the respective video segment i is a facial video segment in the facial video verification case or on whether the facial video segment depicts a talking or a non-talking human face in the case of V-VAD, respectively. In ELM-based classification schemes, the network input weights $\mathbf{W}_{in} \in \mathbb{R}^{K \times L}$ and the hidden layer bias values $\mathbf{b} \in \mathbb{R}^L$ are randomly assigned, while the network output weight $\mathbf{w} \in \mathbb{R}^L$ is analytically calculated. Let us denote by $\mathbf{v}_j$ and w_j the j-th column of $\mathbf{W}_{in}$ and the j-th element of $\mathbf{w}$, respectively. For an activation function $\Phi(\cdot)$, the output o_i of the SLFN network corresponding to the training facial vector $\mathbf{s}_i$ is calculated by:

$$o_i = \sum_{j=1}^{L} w_j \, \Phi(\mathbf{v}_j, b_j, \mathbf{s}_i). \tag{1}$$

It has been shown [34], [35] that almost any nonlinear piecewise continuous activation functions $\Phi(\cdot)$ can be used for the calculation of the network hidden layer outputs, e.g., the sigmoid, sine, Gaussian, hard-limiting and Radial Basis Functions (RBF), Fourier series, etc. In our experiments, we have employed the $RBF - \chi^2$ activation function, which has been found to outperform other choices for BoW-based action classification [36].

By storing the network hidden layer outputs corresponding to the training facial vectors $\mathbf{s}_i$, $i = 1, \ldots, N$ in a matrix $\mathbf{\Phi}$:

$$\mathbf{\Phi} = \begin{bmatrix} \Phi(\mathbf{v}_1, b_1, \mathbf{s}_1) & \cdots & \Phi(\mathbf{v}_1, b_1, \mathbf{s}_N) \\ \cdots & \ddots & \cdots \\ \Phi(\mathbf{v}_L, b_L, \mathbf{s}_1) & \cdots & \Phi(\mathbf{v}_L, b_L, \mathbf{s}_N) \end{bmatrix}, \tag{2}$$

equation (1) can be expressed in a matrix form as $\mathbf{o} = \mathbf{\Phi}^T \mathbf{w}$.

In order to increase robustness to noisy data, by allowing small training errors, the network output weight $\mathbf{w}$ can be obtained by solving for:

$$\textbf{Minimize:} \quad \mathcal{J} = \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{c}{2}\sum_{i=1}^{N} \|\xi_i\|_2^2 \tag{3}$$

$$\textbf{Subject to:} \quad \mathbf{w}^T \boldsymbol{\phi}_i = t_i - \xi_i, \;\; i = 1, ..., N, \tag{4}$$

where ξ_i is the error corresponding to training facial vector $\mathbf{s}_i$, $\boldsymbol{\phi}_i$ is the i-th column of $\mathbf{\Phi}$ denoting the $\mathbf{s}_i$ representation in the ELM space and c is a parameter denoting the importance of the training error in the optimization problem. The optimal value of parameter c is determined by applying a line search strategy using cross-validation. The network output weight $\mathbf{w}$ is finally obtained by:

$$\mathbf{w} = \mathbf{\Phi}\left(\mathbf{K} + \frac{1}{c}\mathbf{I}\right)^{-1} \mathbf{t}, \tag{5}$$

where $\mathbf{K} \in \mathbb{R}^{N \times N}$ is the *ELM kernel matrix*, having elements equal to $[\mathbf{K}]_{i,j} = \boldsymbol{\phi}_i^T \boldsymbol{\phi}_j$ [16], [37].

By using (5), the network response o_l for a test vector $\mathbf{x}_l \in \mathbb{R}^D$ is given by:

$$o_l = \mathbf{W}_{out}^T \boldsymbol{\phi}_l = \mathbf{T}\left(\mathbf{\Phi}^T\mathbf{\Phi} + \frac{1}{c}\mathbf{I}\right)^{-1} \mathbf{k}_l, \tag{6}$$

where $\mathbf{k}_l \in \mathbb{R}^N$ is a vector having its elements equal to $\mathbf{k}_{l,i} = \boldsymbol{\phi}_i^T \boldsymbol{\phi}_l$.

The $RBF - \chi^2$ similarity metric provides the state-of-the-art performance for BoW-based video representations [36], [38]. Therefore, $RBF - \chi^2$ kernel function is used in our experiments:

$$K(i,j) = exp\left(-\frac{1}{4A}\sum_{k=1}^{K}\frac{(s_{ik} - s_{jk})^2}{s_{ik} + s_{jk}}\right), \tag{7}$$

where the value A is set equal to the mean χ^2 distance between the training data $\mathbf{s}_i$.

In order to employ the Dense Trajectory-based facial video representation to train the kernel ELM network described above, a multi-channel kernel learning approach [39] is followed, where:

$$K(i,j) = exp\left(-\sum_{d=1}^{D}\left(\frac{1}{4A}\sum_{k=1}^{K}\frac{(s_{ik}^d - s_{jk}^d)^2}{s_{ik}^d + s_{jk}^d}\right)\right). \tag{8}$$

In most applications where ELM-based classification is performed, classification decision is made solely based on the sign of o_t. However, due to the fact that high precision values, i.e., high true positive rate, are mainly of interest here, a threshold α was introduced in the training phase and fine tuning was performed in order to identify the threshold value giving the best classification precision values.

E. Facial video segment classification (test phase)

In the test phase, a test facial video segment is introduced to the SLFN network. When the STIP-based facial video segment representation is employed, HOG and HOF descriptors are calculated on STIP video locations, L_2 normalized and concatenated, in order to form the corresponding HOG/HOF feature vectors $\mathbf{p}_{tj} \in \mathbb{R}^D$, $j = 1, \ldots, N_t$. $\mathbf{p}_{tj}$ are quantized by using the codebook vectors $\mathbf{v}_k \in \mathbb{R}^D$, $k = 1, \ldots, K$ determined in the training phase and L_1 normalized, in order to produce the facial vector $\mathbf{s}_t$. $\mathbf{s}_t$ is subsequently introduced to the trained kernel ELM network using (7) and its responses o_t are obtained. Similarly, when the Dense Trajectory-based facial video representation is employed, HOG, HOF, MBHx, MBHy, and Trajectory descriptors are calculated on the trajectories of

densely-sampled video frame interest points and $D = 5$ BoW-based video representations $\mathbf{s}_t^d$, $d = 1, \ldots, D$ are produced. $\mathbf{s}_t^d$ are subsequently introduced to the trained kernel ELM network using (8) and its responses o_t are obtained. Finally, the test facial video is classified to the visually talking class if $o_t \geq \alpha$, or to the visually non-talking class if $o_t < \alpha$. In facial video segment verification testing, feature vectors consisting solely of HOG descriptors are also used, both with STIP and with Dense Trajectory-based video segment representation.

In facial video segment verification testing, feature vectors consisting solely of HOG descriptors are also used, both with STIP and with Dense Trajectory-based video segment representation.

IV. Experiments

In this section, experiments conducted in order to evaluate the performance of the proposed approach on V-VAD are presented. One publicly available data set, namely CUAVE as well as a new movie data set containing visual voice activity samples in the wild, were used to this end. A short description of these data sets is provided in the following subsections. Experimental results obtained by SVM and ELM-based classification are subsequently given. Regarding the optimal parameter values used in our method, they have been determined by applying a grid search strategy using the values $c = 10^r$, $r = -6, \ldots, 6$ and $\alpha = 0.1e$, $e = 0, \ldots, 5$.

The classification performance metrics adopted for the evaluation of the classification results achieved by the various methods are classification accuracy (CA), precision (P), F1 measure (F1), miss rate (MR), false acceptance rate (FAR) and half total error rate (HTER = FAR + MAR/2). Moreover, it should be clear by now that, in case no or very slight motion is encountered in a facial video, the adopted video description techniques detect no points of interest, and as a consequence, calculate no descriptors. Even though these videos are omitted during classification, they are taken into consideration in the calculations of the aforementioned performance metrics in the evaluation phase as we make the assumption that they depict either visually silent facial videos or background images which are considered to belong to the visually silent class, too.

A. CUAVE data set

CUAVE [17] is a speaker-independent data set which can be used for voice activity detection, lip reading and speaker identification. It consists of videos of 36 speakers, recorded both individually and in pairs uttering isolated and connected digits standing still in front of a simplistic background of solid color, or slightly moving. The participants are both male and female, with different skin complexions, accents and facial attributes, as can be seen in Figure 4. The facial videos used in our experiments were extracted at a resolution of 195×315 pixels.

Experiments on this data set are usually conducted by performing multiple training-test rounds (sub-experiments), omitting a small percentage of the speakers and using 80% of the remaining for training and the rest 20% for testing, as suggested in [23], [24] and adopted in our experiments. The performance of the evaluated method is subsequently measured by reporting the mean classification rate over all sub-experiments.

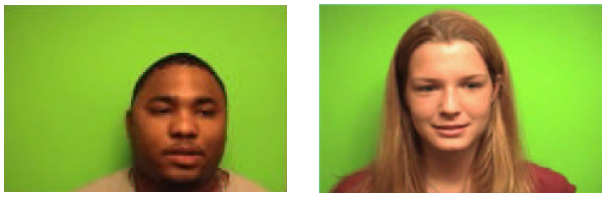

Fig. 4. Sample speakers of the CUAVE data set.

B. Movie data set

The motive for the construction of a data set consisting of facial image videos extracted from full-length movies, was the absence of a data set suitable for (audio)-visual voice activity detection, speech recognition or speaker identification, in the wild (i.e., resembling real-life conditions), as the vast majority of the currently available public data sets are recorded in constrained conditions, e.g., with participants usually standing still in front of a plain background uttering digits, letters, or small phrases. Our data set was, thus, constructed after performing automatic face detection and tracking [11], [12], in three full-length movies. The detected ROIs containing facial images were cropped and resized to fixed size facial images of 195×315 pixels. In some initial exploratory experiments such a resolution was proven adequate for this particular problem. In this way, 4194 video sequences depicting facial image trajectories of 126 actors were extracted in a fully automated way, consisting of facial videos of people of different ages, gender and maybe origin appearing at random poses performing unconstrained movements and talking normally. Moreover, indoor, as well as outdoor shots are encountered, with both stationary and moving complicated backgrounds.

In order for the proposed method to be evaluated on this data set, the leave-one-movie-out cross-validation protocol was applied. Thus, mean classification accuracy results are reported. It should be noted here that, due to the fact that the face detection and tracking were fully automated, some video sequences not depicting facial images also emerged. However, such videos should not exist in a data set oriented for testing V-VAD methods and thus were removed from the data set. This removal can be done either manually or in an automated way. The automatic approach entails the addition of another classification step, prior to the V-VAD step. In this step, the videos are classified based on the presence or absence of human faces in them, using the method described in Section III. Only those classified as facial image videos are fed to the second layer of classifiers, in order to be classified as visually speaking or silent. This preliminary classification step was performed both using all the descriptor histograms calculated for visual speech/silence classification, and utilizing only HOG histograms.

C. Experimental Results

The proposed method has been applied on the CUAVE data set by using the experimental protocols suggested in [23],

[24] after a preprocessing step, which was necessary in order to get frame based results by the proposed method, which normally conducts video based classification. Specifically, a sliding window of length equal to 7 frames moving with step equal to 1 frame was applied on the original videos, in order to split them in smaller parts and labels were assigned to the resulting videos using majority voting on the labels of the frames constituting them. Frame based classification was thus performed, as in [23], [24]. The sliding window length, was chosen in such a way that the number of frames used in V-VAD by the proposed method was equal to the number of frames used for the calculation of the dynamic features exploited by methods [23], [24] for the same purpose.

Table I summarizes in terms of classification accuracy (CA) and visually talking class precision (P) the performance obtained for each experimental setup and each video description method by the aforementioned classification algorithms. As can be seen in this Table, satisfactory visual voice activity detection performance is obtained by applying the proposed method. In more details, the STIP-based video description seems to be more suitable for this data set than Dense Trajectory-based description (DT), achieving better classification accuracies by approximately 15% in both experiments. This can be explained, by taking into account that the combination scheme derived from the second video description method is very complicated, while the visual data set is quite simplistic, thus leading to overtraining and poor generalization in testing.

TABLE I
CLASSIFICATION RATES AND TALKING CLASS PRECISION ON THE CUAVE DATA SET.

CUAVE DS		Experiment [23]		Experiment [24]	
		CA	P	CA	P
STIPs	SVM	87.2%	**87.4%**	86.7%	88.0%
	ELM	**87.6%**	87.0%	**86.8%**	**88.9%**
DT	SVM	74.2%	76.7%	71.4%	73.7%
	ELM	73.8%	75.7%	70.3%	72.4%

Comparison results with other state-of-the-art methods evaluating their performance on the CUAVE data set, are provided in Table II. As can be seen, the proposed method outperforms the classification accuracy of the methods reported in [23], [24] by 15.9% and 12.7%, respectively, on the two experimental setups used in the CUAVE data set, thus achieving great generalization ability on new data. Moreover, in both experiments the proposed method has significantly lower error rates while method [21] seems to be unable to handle the problem posed by this data set.

The results obtained after applying the proposed method on the new, fully unconstrained data set without removing the videos which do not depict facial images are presented in Table III. Satisfactory performance is achieved by both description methods, with a half error rate (HTER) of approximately 30%, that is comparable to the respective performance obtained by state-of-the-art in constrained visual data sets. In addition, dense trajectory-based approach outperforms the STIP-based in all the reported metrics, contrary to what was the case in the CUAVE data set. This can be explained by the fact that in our data set, head movements as well as complex background are encountered. Thus, the descriptors calculated using the dense trajectories method seem to be more efficient, enabling good estimation of face contour and its distinctive motion from that of the background, resulting in better classification rates than those obtained using STIP points description.

The problem whose results were reported on Table III was not the usual V-VAD one, since a third class of samples was also present in the data set, consisting basically of noise. In order to test our method in the real V-VAD problem, we manually removed all the irrelevant videos and performed the experiments again. The results on the "clear" data set are presented in Table IV. By comparing the reported results with those in Table III, a fall in performance metrics rates is noticed in Table IV, especially in the visual silence class, emanating from the removal of irrelevant videos, which were correctly classified as visually silent cases in the previous experiment.

Mean classification results obtained on the three full-length movies constituting the constructed data set, detailed in Section IV-B, are presented in Table V. As can be seen, the facial video segment verification step performs quite well. Very low miss rates are obtained using STIPs and the face class precision as well as the the overall accuracy are satisfactory. Even better results are obtained using Dense Trajectory based description and representation, reaching 93% precision rate, thus allowing the use of this step in the construction of the fully automatic system proposed in this paper, even though the miss rates are slightly worse ($\sim 2-4\%$) than those reported for STIPs.

TABLE V
FACIAL VIDEO SEGMENT VERIFICATION RATES ON THE FULL MOVIE DATA SET.

MOVIE DS		CA	P	MR	F1
STIPs	KSVM	83.6%	85.8%	3.4%	90.8%
	HOG KSVM	**84.0%**	85.0%	**1.7%**	**91.2%**
	KELM	83.8%	**86.5%**	4.2%	90.9%
	HOG KELM	83.8%	86.1%	3.8%	90.8%
DT	KSVM	**94.8%**	91.0%	**5.2%**	92.8%
	HOG KSVM	88.1%	91.5%	5.8%	92.8%
	KELM	89.1%	**93.0%**	6.3%	**93.3%**
	HOG KELM	87.7%	92.1%	7.0%	92.5%

Table VI summarizes the classification results obtained by all the classifier pairs adopted for the automatic removal of non-facial videos from the data set and the subsequent classification of the facial videos as visually speaking and non-speaking. According to them, our approach performs very well, even in the wild, as the classification rates reported are similar to those obtained by other already existing methods on the several simplistic data sets available. Moreover, as already mentioned, STIP-based facial video description is proven inadequate for classification purposes in this case, leading to $\sim 10\%$ lower precision rates and $\sim 5\%$ higher HTER rates than the Dense Trajectory-based method. However, a universal choice of one of the classifier pairs, reported as the best one, would not be right, as depending on the application, different performance metrics are considered the most important. By taking this into consideration, the combination of two neural network based classification steps using Dense-

TABLE II
COMPARISON RESULTS ON THE CUAVE DATA SET.

CUAVE DS	Experiment [23]				Experiment [24]			
	CA	HTER	FAR	MR	CA	HTER	FAR	MR
Method [21]	52.8%	47.1%	40.8%	53.3%	52.6%	47.2%	41.0%	53.5%
Method [23]	71.3%	25.6%	31.8%	28.7%	-	-	-	-
Method [24]	-	-	-	-	74.1%	25.9%	24.2%	27.6%
Proposed method	**87.2%**	**11.3%**	**14.1%**	**8.5%**	**86.8%**	**11.4%**	**11.5%**	**11.3%**

TABLE III
CLASSIFICATION RATES ON THE FULL MOVIE DATA SET.

CONSTRUCTED DS	Full data set		Visual silence			Visual speech		
	CA	HTER	P	FAR	F1	P	MR	F1
STIPs	70.8%	37.7%	71.8%	8.9%	80.2%	68.6%	66.4%	44.0%
DT	**76.4%**	**30.5%**	**76.1%**	**7.3%**	**83.6%**	**77.6%**	**53.8%**	**57.9%**

TABLE IV
CLASSIFICATION RATES ON THE "CLEAR" MOVIE DATA SET.

CONSTRUCTED DS	Full data set		Visual silence			Visual speech		
	CA	HTER	P	FAR	F1	P	MR	F1
STIPs	67.8%	35.5%	68.5%	15.4%	75.5%	67.8%	55.6%	52.8%
DT	**71.1%**	31.3%	**69.9%**	13.2%	77.2%	**74.8%**	49.4%	60.3%

Trajectory based facial video description can be regarded as the best alternative. This is in line with the remark that in our experiments, we mainly focus on the minimization of false detection error, and thus, on the maximization of visually speaking class precision metric.

TABLE VI
CLASSIFICATION RATES ON THE AUTOMATICALLY CLEARED MOVIE DATA SET.

MOVIE DS		CA	HTER	P
STIPs	KSVM-KSVM	68.5%	37.0%	62.2%
	HOG KSVM-KSVM	70.9%	35.9%	67.5%
	KSVM-KELM	69.7%	37.8%	68.2%
	HOG KSVM-KELM	**70.8%**	36.7%	**68.2%**
	KELM-KSVM	70.1%	36.4%	67.3%
	HOG KELM-KSVM	70.7%	**35.8%**	67.5%
	KELM-KELM	69.3%	37.3%	64.9%
	HOG KELM-KELM	69.6%	37.2%	65.8%
DT	KSVM-KSVM	73.0%	29.8%	70.9%
	HOG KSVM-KSVM	73.0%	**29.6%**	71.2%
	KSVM-KELM	73.1%	31.0%	76.5%
	HOG KSVM-KELM	73.2%	30.7%	77.5%
	KELM-KSVM	72.5%	29.7%	71.1%
	HOG KELM-KSVM	72.6%	29.8%	71.0%
	KELM-KELM	73.2%	30.3%	**78.8%**
	HOG KELM-KELM	**73.4%**	30.3%	78.6%

Finally, based on the results reported in Table VII, our method is proven to be much more efficient than one of the current state-of-the-art methods for visual voice activity detection, as it outperforms it by 23.8%. More specifically, method [21] which was tested only on facial videos of frontal images, seems to fail in dealing with the unconstrained problem, while the proposed method achieves satisfactory classification accuracy. The poor performance of the method [21] in this data set was to a great extend expected, as its implementation utilizes face proportions in order to perform mouth detection. This approach is successfully applicable only in frontal facial images and apparently fails in cases, where face rotation of more than $\sim 30^\circ$ horizontally and/or $\sim 10^\circ$ vertically are encountered, which are very frequent in our data set.

V. CONCLUSIONS

In this paper, we proposed a novel method for Visual Voice Activity Detection in the wild that exploits local shape and motion information appearing at spatiotemporal locations of interest for facial video description and the BoW model for facial video representation. SVM and Neural Network-based classification based on the ELM using the BoW-based facial video representations leads to satisfactory classification performance. Experimental results on one publicly available data set denote the effectiveness of the proposed method, since it outperforms recently proposed state-of-the-art methods in a user independent experimental setting. The respective results on the fully unconstrained data of a new movie data set especially constructed for dealing with the V-VAD problem in wild, prove the efficiency of the proposed method even in the unconstrained problem, in which state-of-the-art methods fail.

ACKNOWLEDGEMENT

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 287674 (3DTVS). This publication reflects only the author's views. The European Union is not liable for any use that may be made of the information contained therein.

REFERENCES

[1] G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1254–1265, November 2009.

TABLE VII
COMPARISON RESULTS ON THE CONSTRUCTED DATA SET.

CONSTRUCTED DS	Full data set		Visual silence			Visual speech		
	CA	HTER	P	FAR	F1	P	MR	F1
Method [21]	49.6%	49.2%	57.2%	64.9%	43.1%	45.2%	33.5%	53.8%
Proposed method	**73.4%**	**30.3%**	**71.5%**	**9.3%**	**80.0%**	**78.6%**	**51.4%**	**60.0%**

[2] C. Zhang, P. Yin, Y. Rui, R. Cutler, P. Viola, X. Sun, N. Pinto, and Z. Zhang, "Boosting-based multimodal speaker detection for distributed meeting videos," *IEEE Transactions on Multimedia*, vol. 10, no. 8, pp. 1541–1552, December 2008.

[3] K. Nathwani, P. Pandit, and R. Hegde, "Group delay based methods for speaker segregation and its application in multimedia information retrieval," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1326–1339, October 2013.

[4] M. Sargin, Y. Yemez, E. Erzin, and A. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Transactions on Multimedia*, vol. 9, no. 7, pp. 1520–1403, November 2007.

[5] Q. Liu, A. Aubrey, and W. Wang, "Interference reduction in reverberant speech separation with visual voice activity detection," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1610–1623, October 2014.

[6] V. Minotto, C. Jung, and B. Lee, "Simultaneous-speaker voice activity detection and localization using mid-fusion of SVM and HMMs," *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 1032–1044, June 2014.

[7] S. Petridis and M. Pantic, "Audiovisual discrimination between speech and laughter: Why and when visual information might help," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 216–234, April 2011.

[8] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2–3, pp. 107–123, September 2005.

[9] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," *Computer Vision and Pattern Recognition*, pp. 3169–3176, 2011.

[10] A. Iosifidis, A. Tefas, and I. Pitas, "Minimum class variance extreme learning machine for human action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 11, pp. 1968–1979, November 2013.

[11] G. Stamou, M. Krinidis, N. Nikolaidis, and I. Pitas, "A monocular system for person tracking: Implementation and testing," *Journal on Multimodal User Interfaces*, vol. 1, no. 2, pp. 31 – 47, 2007.

[12] O. Zoidi, A. Tefas, and I. Pitas, "Visual object tracking based on local steering kernels and color histograms," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 5, pp. 870 – 882, 2013.

[13] Y. Huang, Z. Wu, L. Wang, and T. Tan, "Feature coding in image classification: A comprehensive study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 493–506, 2014.

[14] A. Iosifidis, A. Tefas, and I. Pitas, "Discriminant bag of words based representation for human action recognition," *Pattern Recognition Letters*, vol. 49, pp. 185–192, 2014.

[15] G. Huang, Q. Zhu, and C. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks," *International Joint Conference on Neural Networks*, vol. 2, pp. 985–990, July 2004.

[16] A. Iosifidis, A. Tefas, and I. Pitas, "On the kernel extreme learning machine classifier," *Pattern Recognition Letters, D.O.I. 10.1016/j.patrec.2014.12.003*, 2014.

[17] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," *International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. II–2017 – II–2020, May 2002.

[18] S. Takeuchi, H. Takashi, S. Tamura, and S. Hayamizu, "Voice activity detection based on fusion of audio and visual information," *AVSP*, pp. 151–154, 2009.

[19] D. Sodoyer, B. Rivet, L. Girin, J.-L. Schwartz, and C. Jutten, "An analysis of visual speech information applied to voice activity detection," *International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. I–I, 2006.

[20] D. Sodoyer, B. Rivet, L. Girin, C. Savariaux, J.-L. Schwartz, and C. Jutten, "A study of lip movements during spontaneous dialog and its application to voice activity detection," *The Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. 1184–1196, 2009.

[21] S. Siatras, N. Nikolaidis, and I. Pitas, "Visual speech detection using mouth region intensities," *European Signal Processing Conference*, 2006.

[22] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, and T. Darrell, "Visual speech recognition with loosely synchronized feature streams," *International Conference on Computer Vision*, vol. 2, pp. 1424–1431, 2005.

[23] R. Navarathna, D. Dean, P. Lucey, S. Sridharan, and C. Fookes, "Dynamic visual features for visual-speech activity detection," *Conference of International Speech Communication Association*, 2010.

[24] R. Navarathna, D. Dean, S. Sridharan, C. Fookes, and P. Lucey, "Visual voice activity detection using frontal versus profile views," *International Conference on Digital Image Computing Techniques and Applications*, pp. 134–139, 2011.

[25] Q. Liu, W. Wang, and P. Jackson, "A visual voice activity detection method with adaboosting," *Sensor Signal Processing for Defence (SSPD 2011)*, pp. 1–5, 2011.

[26] A. Aubrey, Y. Hicks, and J. Chambers, "Visual voice activity detection with optical flow," *IET Image Processing,*, vol. 4, no. 6, pp. 463–472, 2010.

[27] I. Almajai and B. Milner, "Using audio-visual features for robust voice activity detection in clean and noisy speech," *European Signal Processing Conference*, vol. 86, 2008.

[28] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421 – 2424, November 2006.

[29] V. Minotto, C. Lopes, J. Scharcanski, C. Jung, and B. Lee, "Audiovisual voice activity detection based on microphone arrays and color information," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 147–156, 2013.

[30] I. Laptev and T. Lindeberg, "Space-time interest points," *Internationa Conference on Computer Vision*, pp. 432–439, 2003.

[31] C. Harris and M. Stephens, "A combined corner and edge detector," *Alvey Vision Conference*, pp. 147–152, 1988.

[32] H. Wang, M. Ullah, A. Kläserr, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," *British Machine Vision Conference*, 2009.

[33] S. Theodoridis and K. Koutroumbas, "Pattern recognition," *Academic Press*, 2008.

[34] G. B. Huang, L. Chen, and C. K. Siew, "Universal approximation using incremental constructive feedforward networks with random hidden nodes," *IEEE Transactions on Neural Networks*, vol. 17, no. 4, pp. 879–892, 2006.

[35] G. B. Huang and L. Chen, "Convex incremental extreme learning machine," *Neurocomputing*, vol. 70, no. 16, pp. 3056–3062, 2008.

[36] A. Iosifidis, A. Tefas, and I. Pitas, "Minimum variance extreme learning machine for human action recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5427–5431, 2014.

[37] G. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 2, pp. 513–529, 2012.

[38] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," *British Machine Vision Conference*, 2009.

[39] J. Zhang, M. Marszalek, M. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.